Graphical Phone Representation

Amr M. Gody*
Cairo University

The problem of phone recognition is transferred to the visual domain. Each phone is represented by unique graphical Character. The work is applied on a single speaker to eliminate the effect of multi speaker in this phase. The system was trained using hand-verified data from single speaker. Using 31 context-dependent phone models, a baseline phone accuracy of 63.5% (no phone grammar) has been obtained on an independent test set segments from the same speaker. The evaluation is based on human eye recognition of the generated visual characters.

1. Introduction

Phone recognition is a very difficult task in speech recognition. Although phone is the smallest speech unit, no commercial recognizer is built based on the phones. All commercial recognizers are built on word recognition using statistical models such as HMM. Thousands of words are required to train the system. And tens of rules are applied to verify the system.

Many researches are done in the way to get a robust phone recognition system. In 1991, AT&T Bell Labs introduces an automatic approach to segmentation of labeled speech by Ljolje, A. and Riley, M.D.[1]. The authors investigate an automatic approach to segmentation of labeled speech and labeling and segmentation of speech when only the orthographic transcription of speech is available. The technique is based on a phone recognition system based on a trigram phonotactic model, gamma distribution phone duration models, and a spectral model based on five different structures for phone models of varying contextual dependencies. The alignment of speech with a given phone sequence is performed as a very constrained phone recognition task with the phonotactic model based only on the given phone sequence. When only orthographic transcription is provided, a classification-tree-based prediction of most likely phone realizations is used as an input network for the phone recognizer. The maximum likelihood phone sequence is then treated as the true phone sequence and its segment boundaries are compared with the reference boundaries

In 1992 Lamel, L.F. and Gauvain, J.-L. executed a series of experiments for speaker-independent continuous speech phone recognition [2]. The authors' experiments were the first to use BREF corpus database, and were meant to provide a baseline performance evaluation for vocabulary independent phone recognition. The system was trained using hand-verified data from 43 speakers. Using 35 context-dependent phone models, a baseline phone accuracy of 60% (no phone grammar) has been obtained on an independent test set of 7635 phone segments from 19 speakers. Including phone bigram probabilities as phonotactic constraints results in a performance of 63.3%. A phone accuracy of 68.6% (73.3% correct) was obtained with 428 context dependent models.

* Department of Electrical Engineering, Faculty of Engineering, Cairo University – Fayoum Branch, El-fayoum , EGYPT., Email: agody@ieee.org.

In 1994, the department of Engineering, Cambridge University introduced a work in Phone modeling [3]. The work was implemented using recurrent neural networks RNN. The work described phone-modeling improvements to the hybrid connectionist-hidden Markov model speech recognition system developed at Cambridge University. These improvements were applied to phone recognition from the TIMIT task and word recognition from the Wall Street Journal (WSJ) task. A recurrent net was used to map acoustic vectors to posterior probabilities of phone classes. The maximum likelihood phone or word string was then extracted using Markov models. The paper described three improvements: connectionist model merging; explicit presentation of acoustic context; and improved duration modeling.

In 1996, the school of Electrical & computer engineering, Purdue University introduced a work in implementation of neural network in speech phone recognition [4]. Seven different criterion functions were evaluated for speech recognition. A new criterion function that allows direct minimization of the frame error rate was proposed. Two new optimization methods for RNN weight updating were investigated. Experiments have been carried out on the Intel Paragon parallel processing system. They didn't mention the rate of success explicitly but they mentioned that the results were competitive with the best results in the literature.

In 1996, the school of Electrical & computer engineering, Purdue University presented a new statistical speech model in which coarticulation is modeled explicitly [5]. Unlike HMMs, in which the current state depends only on the previous state and the current observation, the proposed model supports dependence on the previous and next states and on the previous and current observations. The degree of coarticulation between adjacent phones is modeled parametrically, and can be adjusted according to a parameter representing the speaking rate. The model also incorporates a parameter that represents a frame-by-frame measure of confidence in the speech. Two methods for solving the system parameters were presented: one based on the K-means method, and a novel method based on explicitly minimizing a measure of the segmentation error. A new, efficient forward algorithm and the use of top candidates in the search greatly reduce the computational complexity. In evaluation on the TIMIT database, phone recognition rate of 77.1% was achieved.

2. Graphical Character Generator GCG

In the presented paper, the goal is to obtain a stable character representation for each phone. The characters verification by human eye is considered as a measure of success mapping. Database is prepared for a single speaker. Multi-speaker effect will be modeled in future work. The elimination of Multi-speaker effect from the generated character is a step toward robust speaker independent phone recognizer. The system used to generate the visual characters is represented in figure1.

A stream of n digitized speech samples is applied to BLK0. A frame of 20 ms is chosen as analysis frame length. No waiting window is applied and no frame overlapping is considered. The output vector F_{1x220} is applied to BLK1 for normalization process. This step is important to eliminate the environmental effect from the speech signal under test. The analysis frame Fn_{1x220} is ready for the analysis phase.

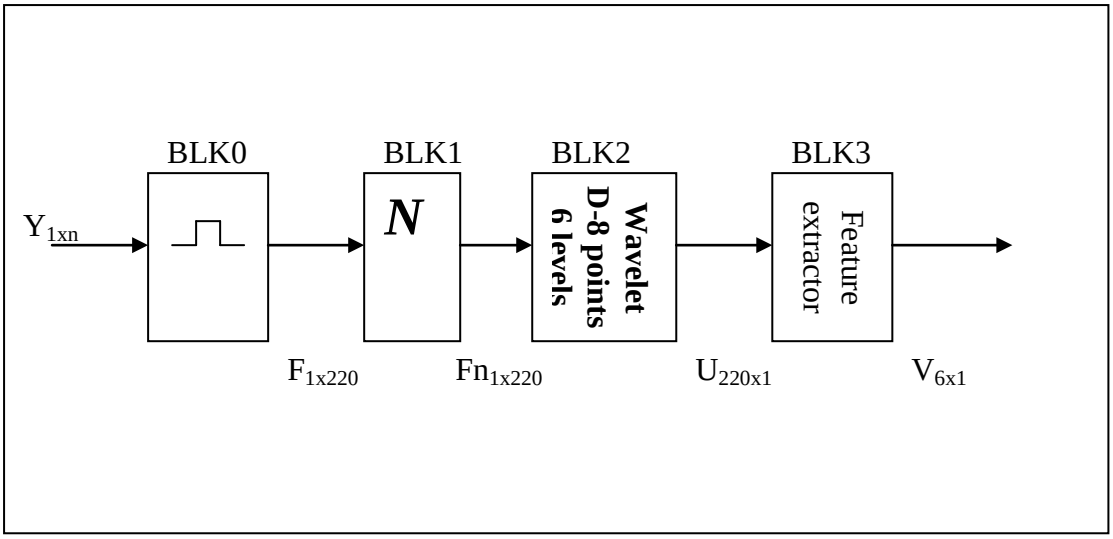


Figure 1. Block diagram of Graphical Character Generator subsystem (GCG).

Fn_{1x220} is introduced to BLK2 to generate frequency domain time projections of the analysis Frame (U_{220x1}). U_{220x1} is a worthy informative vector. It gives an answer of the question "What is the frequency content at time t of this Frame". The wavelet filter is chosen to be Doubechies of 6 levels in depth. This gives the signal frequency projections in 6 different frequency bands. U_{220x1} is applied to BLK3. This block is considered as a feature generator. The features here are the R.M.S values of the signal in each frequency band. This gives us a vector of 6 elements as each element represents the power in the signal in the corresponding frequency band. The vector V_{6x1} is generated from BLK3. Table 1 gives explanation of V_{6x1} vector.

Table 1, vectors explanation for the system considering 11025 (Hz) sampling rate and 220 (ms) analysis frame length[6]

Frequency range (KHz)	Element index in V_{6x1}
2.7-5.5	1
1.4-2.7	2
0.69-1.4	3
0.344-0.69	4
0.172-0.344	5
0.086-0.172	6

As table 1 indicates, the projected signal in each band has a different resolution. An optional stage of interpolation is applied to align the length of the generated projections as the length of the analysis frame (220 samples). This stage is for mathematical interpretation for further steps. It can be eliminated in the final practical system.

Figure 2, represents a graphical drawing of U_{6x1}. The x-axis represents the element index in U_{6x1} and the y-axis represents the R.M.S. value corresponding the index.

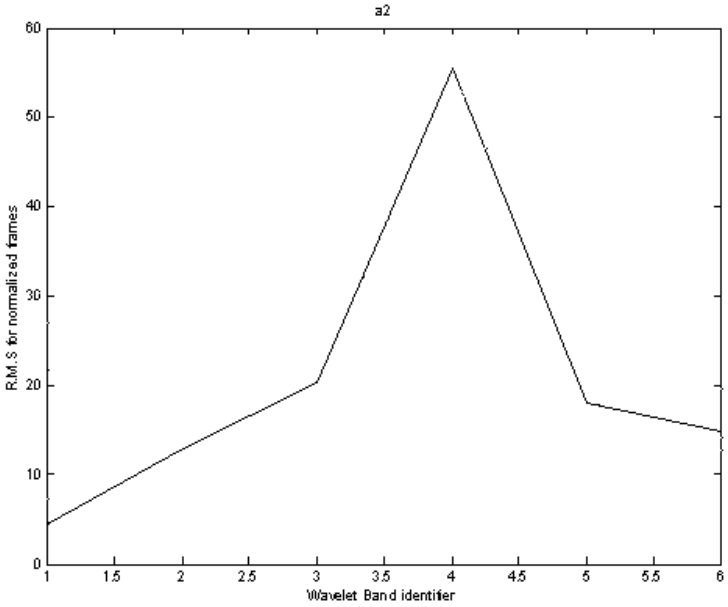


Figure 2. U_{6x1} is drawn for a segment of Arabic vowel /a/.

As shown in figure 2 the character has a peak at 4. Reviewing table 1 indicates that band 4 represents the frequency range from 344-690 (Hz) which is logical for vowel /a/.

Figure 3 indicates how far the character shape is stable over hundreds of segments of vowel /a/ taken from different parts of different words.

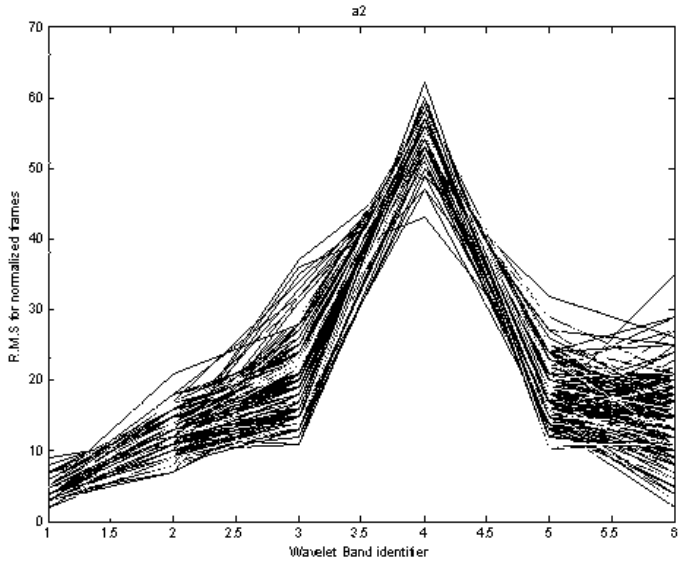
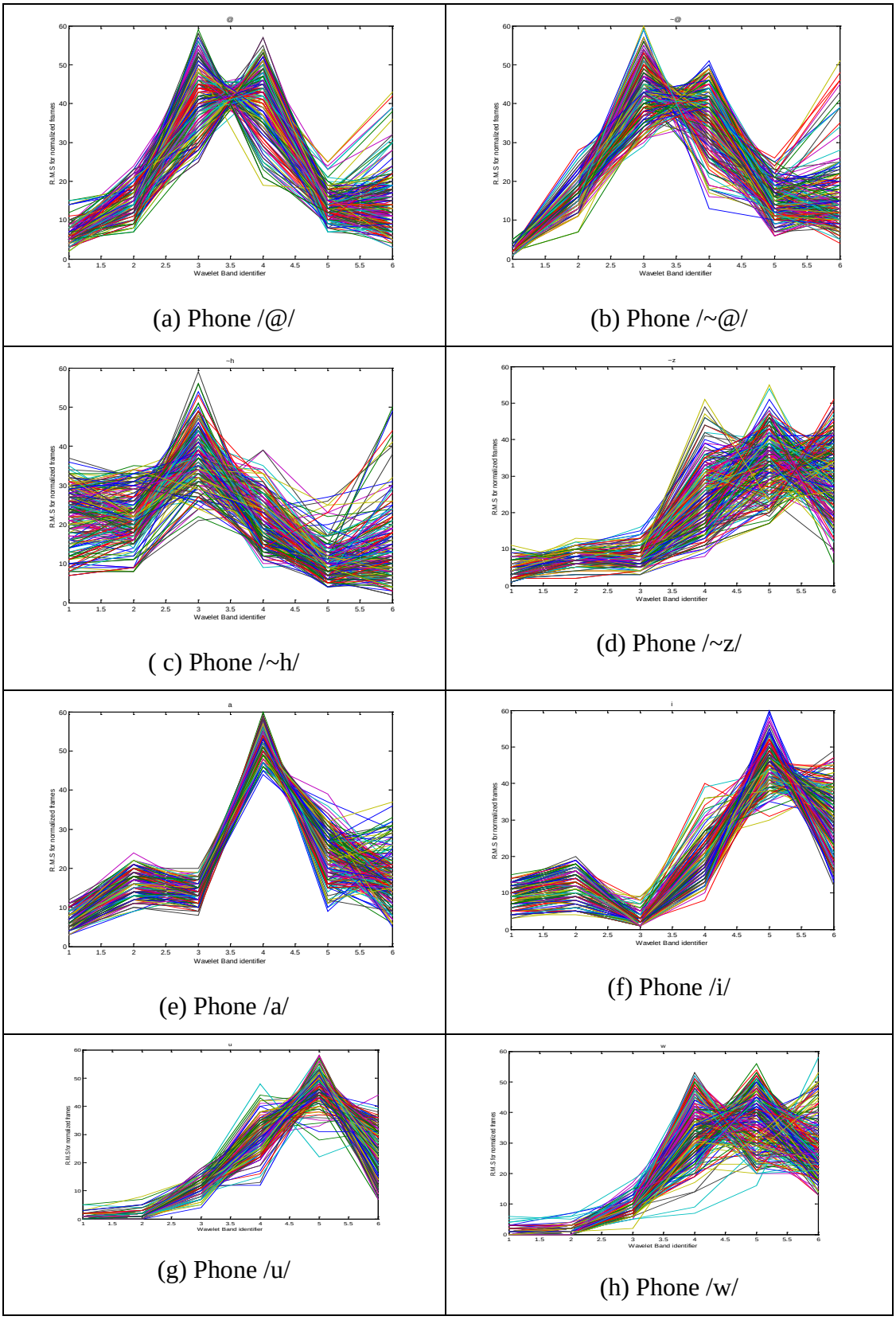


Figure 3. U_{6x1} representation of about 5000 ms total durations of vowel /a/ taken from different parts of different words.

3. The procedure

A database is prepared to get equally-length training stream for each phone in the Arabic language (28 consonants + 3 vowels). The segmentation process is hand-verified to verify phone boundaries. The spoken words are selected so that each word contributes one phone that can be easily detected.

Graphical Character Generator subsystem (CGC) process is applied for each phone-composed stream. Each phone stream is 5000 (ms) in length. Figure 4 illustrates the results of some phones.



(a) Phone /@/

(b) Phone /~@/

(c) Phone /~h/

(d) Phone /~z/

(e) Phone /a/

(f) Phone /i/

(g) Phone /u/

(h) Phone /w/

Figure 4. Some phone characters.

Figure 4, gives a brief description of the goal of this research. As shown in this figure that contains some of the 31 phones (28 consonants +3 vowels), see table 2 for more details, in the Arabic language how they are highly discriminated by eye. We can notice how far the character is stable for the same phone. One of the problems in phone recognition is how we can discriminate the vowels. As shown in figure 4 (e), (f) and (g) the three Arabic vowels are highly discriminated.

4. System evaluation

A new phones stream that represents all Arabic phones is composed for the test process. The system evaluation is based on human eye recognition. The evaluation indicates an average success of 63.5%.

4.1 Test data preparation

A test data is prepared to be ready for comparison with reference pattern process. Test data features are obtained as discussed in section 2. A matrix of 7x1042 is prepared, where 1024 represents the total number of test frames and 7 represents the number of elements in each test vector. The vector is structured such that the first element represents the phone ID and the rest of the elements represent the Character vector to be compared with the reference vector.

Test Vector = [ID, GC elements]

Figure 5 indicates the complete set of Arabic phones under test and table 2 show the symbol keys in Arabic language.

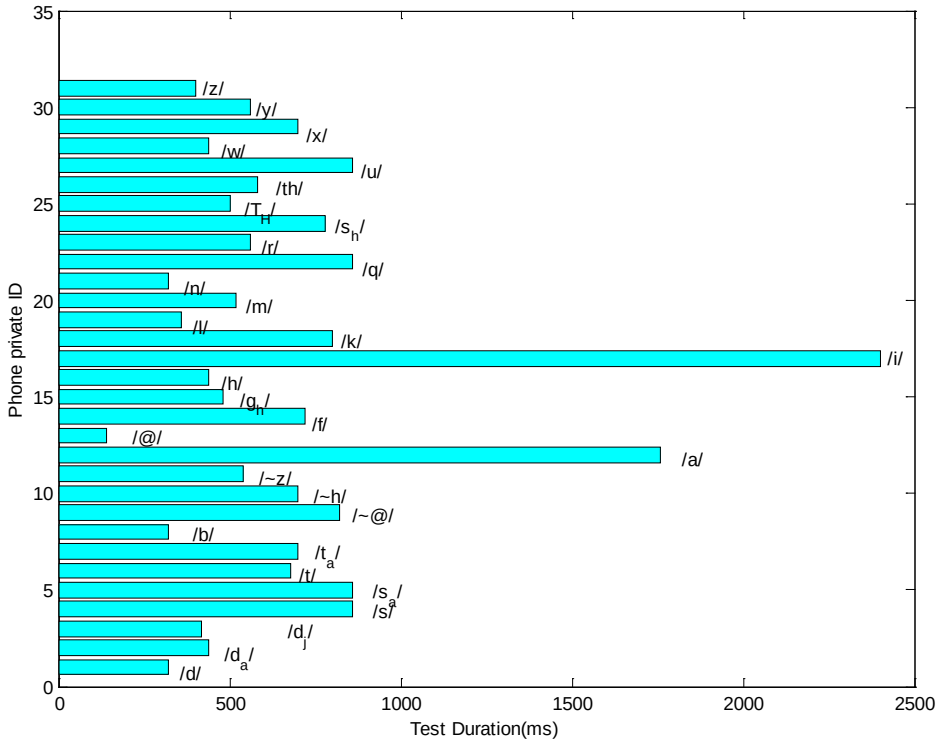


Figure 5. Test set durations

Table 2, symbols key table.

Phone		
#	Context symbol	Speech symbol
1	ء	@
2	ب	B
3	ت	T
4	ث	Th
5	ج	D_J
6	ح	~H
7	خ	X
8	د	D
9	ذ	~Z
10	ر	R
11	ز	Z
12	س	S
13	ش	S_H
14	ص	S_A
15	ض	D_A
16	ط	T_A
17	ظ	T_H
18	ع	~@
19	غ	G_H
20	ف	F
21	ق	Q
22	ك	K
23	ل	L
24	م	M
25	ن	N
26	هـ	H
27	و	W
28	ي	Y
29	ُ	U
30	َ	A
31	ِ	I

4.2 Human eye mathematical model

In this work the focus is to discriminate phones by eye. No rules are added in this stage to get just a vision of how this system is promising. To evaluate such huge data by eye with minimum errors, the eye is modeled mathematically. i.e a mathematical model that simulates the human eye is obtained. The idea is to get what you see model for the generated characters. N vectors represent each character in the training phase. The mean vector is obtained for each character. Also the standard deviation of each character is obtained. If more than 5 points of the test character are plotted within the curves (mean and standard deviation) then the test character belongs to this set. Figure 6 represents the idea. As shown in figure 6, the points marked as 'x' are dropped within the upper and the lower curves so these points represent a character which belongs to this set.

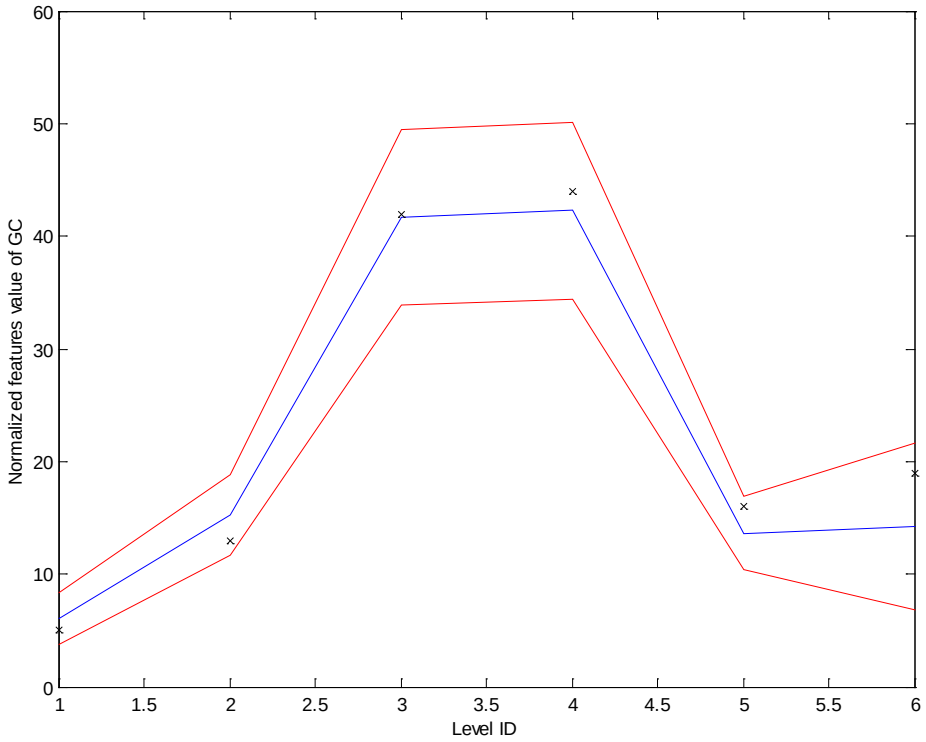


Figure 6, Human eye can see the marked 'x' points as a character which belongs to the drawn set.

This model is a must to get practical results. The human brain can not remember the character shape all the time and it will be so tired after trying 100 samples which is not practical to get practical results. So this model acts as a human eye and is applied in the automation process of evaluation.

4.3 System results

After applying the mathematical human eye a further step is added to get the best performance that can be obtained from this system. A simple weighting process is applied to the vectors under test. The weighting vector consists of 6 digits each one is applied to a point in the vector under test (see figure 5 'x' points). The system results are tabulated for each possible combinations of weight vector. The weight vector acts

as an on off not more in this stage. It is just used to get a vision of the dominant feature elements. So we have 64 combinations of the weight vectors. This means that the experiment will be repeated 64 times to get the best hit. To simplify the process, the combinations that contain more than 3 zeros are excluded. Figure 7 indicates the results of these experiments. As shown in the figure the weight 56 (111000) in binary gives the highest mean recognition accuracy. Actually this result gives an indication that the low frequencies are the dominant in the recognition process. See table 1 that link between the frequency bands and the level ID number. As the LSB represents level 1 and the MSB represents level 6.

Figure 8 gives the best recognition percentage diagram that can be obtained with this simple and direct eye model. No rules or grammars are applied.

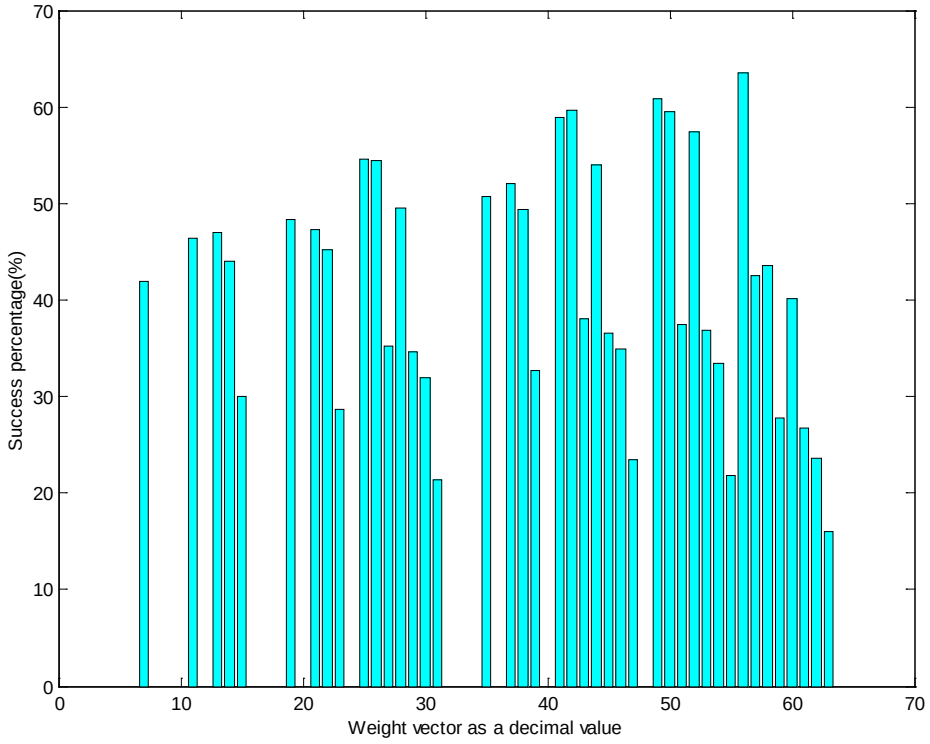


Figure 7. The results of the 64 experiments to obtain the best system performance point

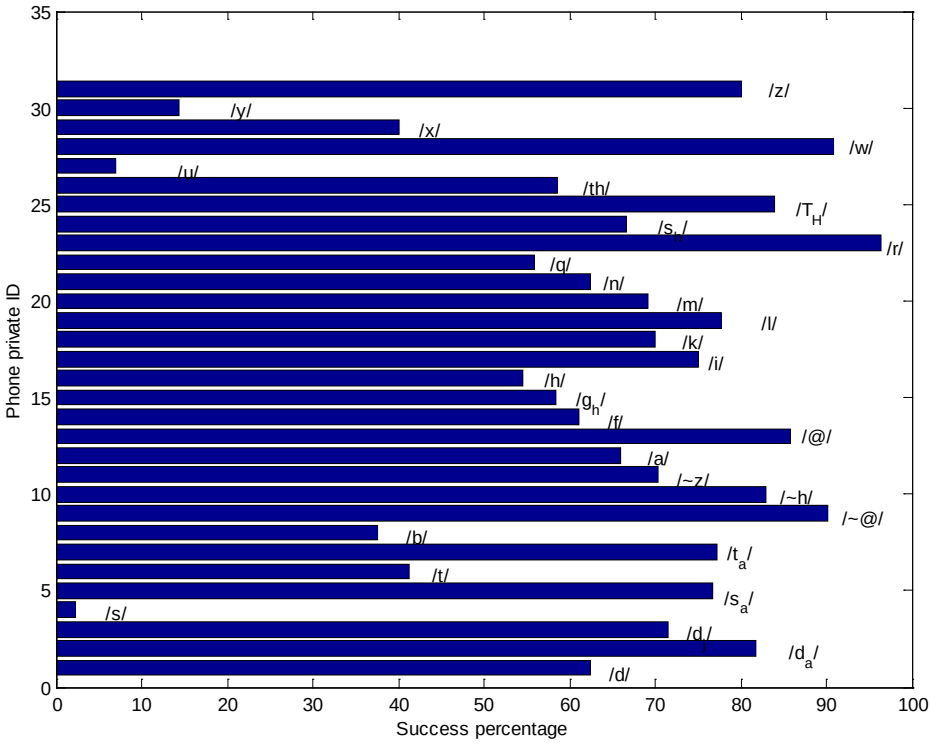


Figure 8. Individual phone recognition percentage

The results can be abbreviated as shown in figure 9.

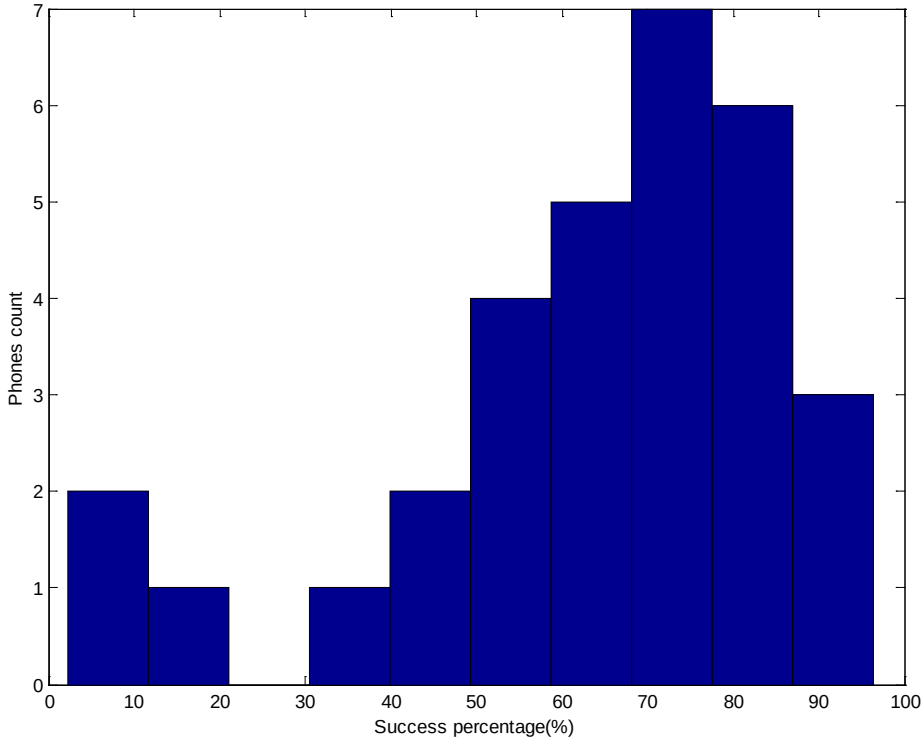


Figure 9. Histogram for the system results.

Figure 9, indicates that about 3 phones gives less than 30% and 21 phones gives more than 60%.

5. Conclusion

Phones can be represented graphically using the proposed method. This work is promising for very robust machine recognition. The problem of phone recognition is totally transferred to image recognition domain. The system is evaluated using human eye recognition based method. It gives an overall efficiency of 63.5%. Multi speaker effect is excluded in this work to simplify the process of mapping as a first step. In future work, this effect will be modeled as a noise embedded into the characters as a step toward the elimination. The procedure of mapping is very simple and can be easily implemented in real time applications. It is promising for Real-Time-Robust-Speaker independent recognition system.

6. References

[1] Ljolje, A.; Riley, M.D.,"Automatic segmentation and labeling of speech", Acoustics, Speech, and Signal Processing, ICASSP-91., 1991 International Conference on , 14-17 Apr 1991,Page(s): 473 -476 vol.1

[2] Lamel, L.F.; Gauvain, J.-L.,"Experiments on speaker-independent phone recognition using BREF", Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on , 23-26 Mar 1992 , vol.1, Page(s): 557 -560

[3] Robinson, T.; Hochberg, M.; Renals, S.,"IPA: improved phone modelling with recurrent neural networks", Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on , 19-22 Apr 1994, vol.1, Page(s): I/37 -I/40

[4] Ruxin Chen; Jamieson L., "Experiments on the implementation of recurrent neural networks for speech phone recognition", Signals, Systems and Computers, 1996. 1996 Conference Record of the Thirtieth Asilomar Conference on , 3-6 Nov 1996, vol.1, Page(s): 779 -782

[5] R. Chen, L. H. Jamieson, "Explicit modeling of coarticulation in a statistical speech recognizer", Proc. Int. Conf. Acoustic, Speech, Signal Processing, Atlanta, GA, Page(s): 463-466, May 1996.

[6] Amr M. Gody, “Natural Hearing Model Based On Dyadic Wavelet”, The third conference on language Engineering CLE’2002, Page(s): 37-43,October 2002